Weekly Report
12/08/2014 - 12/14/2014

Jing Xia

December 14, 2014

1 Summary

This week I mainly focused on the rank project, the data inspection project and the driving class.

2 Projects

2.1 Project 1 - Rank Visualization

Considering the poor network connection, I built up a graph database myself to archive the page link connection. Currently it's still running to fetch all page link relations for all top-1000 pages.
The appendix is the current draft of the paper, just abstract and introduction.

2.2 Project 2 - Data Inspection

I further summarize the ideas for the data visual inspection (DataVIP) system as follows:

- **Layers of view graph** are explored progressively, so is the underlying data, by users selecting a sub-dataset from one view.
- For each layer, the view graph has **a graph layout** that keep similar views close and dissimilar views apart.
- For each view, the view node? is represented with **a glyph-based view descriptor** that characterizes the data distribution and its chart type.

And the major functions to be implemented include:

- definition of correlation - positive correlation, negative correlation and other nonlinear correlations
- view graph layout - that describes correlations of view overview and selection-based view overview

- glyph-based view descriptor - a view descriptor that characterizes the visualization, the underlying data type as well as the data distribution.
- view graph comparison - vertical comparison (comparison between layers), horizontal comparison (comparison within layers)

2.3 Project 3 - NBA Game Visualization

We've got Twitter data from Jiawei. But the dataset for one game is too small (600+ tweets in total), 3-5 tweets per minute. It is not enough in evaluating game process.
We also found a website doing similar game analysis with NBA statistical dataset (http://popcornmachine.net/gf?date=20141209&game=DALMEM). Our current plan is to enhance the data analysis with visualization and to enable game-wise comparison.

3 Paper Reading

-

4 Miscellaneous

This week I had four driving classes.

5 To Do List

1. TopK: adjustment of similarity measurement
2. TopK: A more detailed draft of paper.
3. NBA project discussion and data inspection project discussion.

References

[1] Shi-Sheng Huang, Ariel Shamir, Chao-Hui Shen, Hao Zhang, Alla Sheffer, Shi-Min Hu, and Daniel Cohen-Or. Qualitative organization of collections of shapes via quartet analysis. *ACM Transactions on Graphics*, 32(4):1, July 2013.

[2] Jock Mackinlay, Pat Hanrahan, and Chris Stolte. Show me: Automatic presentation for visual analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1137–1144, 2007.

[3] M Vartak, S Madden, and A Parameswaran. SEEDB: Automatically Generating Query Visualizations. In *Proceedings of the VLDB*, pages –, January 2014.

Appendices

A Abstract

Wikipedia top page view statistics are collections of top viewed Wikipedia pages over time, and of great importance in analyzing users' interest or current affairs. However, visualizing Wikipedia top page-view statistics usually suffers from great visual clutter. We formed a principle that any good design should connect the same Wikipedia page over time without causing unnecessary perceptual complexity. Following gestalt's law of continuity, we tried out a variety of visual designs of top ranked pages. Also, users are able to explore connections among those top viewed pages by taking both the page-view behavior and the page-link information into consideration. Such combination enhances the unweighted Wikipedia page-link network and brings users' page of interest in focus. We conducted a user study of the various visual designs and evaluated the usage of the system. The results show the feasibility of the visual design and the system.

B Introduction

Wikipedia is considered as the biggest online encyclopedia, everyday page-view can be more than 600M in all languages. It has become a knowledge exchange market where users learn and contribute their knowledge. Due to the accumulation, Wikipedia has also become a huge knowledge warehouse.
As opposed to various researches regarding infrastructures of Wikipedia such as NLP and data warehouse, work on time variance of Wikipedia top page-views is still new. According to the 20/80 principle, the ranking data of Wikipedia top page-view statistics (Wikipedia page rank for short) reflects users' major interests in Wikipedia or furthermore in current affairs. Time series of Wikipedia page-rank, or Wikipedia page-ranking trends therefore conveys how users' social interest evolving over time.

C Various Mutants of Page-Rank Representation

In the visualization of page-rank time series, we followed the gestalt law of continuity and had several design proposals. A general goal is to describe the local rank-involving trend of a specific page. We conducted a user study for the best representation of page-rank time series.

C.1 The Sparkline Visualization

An intuitive solution of time series would be sparkline of local rank changes. Users would be able to see a rough local trend of any given page and make

comparisons among sparklines.
However, the drawback is also obvious. It becomes so visually unpleasant when it comes to display full of worm-like sparklines. Also, due to limited display space, subtle changes cannot be accurately told and users do not know what is the exact next rank of the page.

C.2 The Badge Visualization

Guided by the gestalt law of continuity, we get rid of the lines or streams that cause visual clutter but retain the page rank glyph meanwhile.

D Semantic Exploration

We explore the page-wise semantics with two integrated relations: **the page-link relation** and **the rank time series similarity relation**. The page link relation states a semantic relation explicitly, while similarity relation is more implicit. For a current affair, users tend to query several key words that are related to the affair, resulting in concurrent rank time series. Under this assumption, we indicate that those pages with similar rank series are potentially related. However, users should note that it is not necessary that page with similar rank series are linked. We are simply trying to enhance the planar page link network with user behaviors and draw users' limited attention to a more condense network.

D.1 Finding Pages of Similar Trend

We evaluate the dissimilarity between two pages with the proximity of their associated rank time series, under the assumption that two Wikipedia pages are potentially but not necessarily correlated if they share similar trends during their recent history. We first take all rank series as pollens and compare their similarity with curving We define the dissimilarity of two ranked items at a time point as the weighted sum of two factors: the curve matching factor f_{cm} and the loss compensation factor f_{comp}.

$$Dissimilarity = \frac{w_{cm} * f_{cm} + w_{comp} * f_{comp}}{w_{cm} + w_{comp}}$$

The curve matching factor is the main factor for rank series similarity. obtained by a curve matching method [] which applies dynamic time warping in curve matching. The paper takes similarity of two polygonal chains equivalent to the optimal distance on the manifold made by the two chains. The optimal distance can be calculated with dynamic time warping. In addition, to get a more accurate results, it interpolates steiner points on the edges of the manifold patches. To adapt the method we can take rank time series as polylines and compare their similarity with the method described in this paper.

Although f_{cm} considers the influence of discontinuous rank time series, its confidence is low when the underlying time series has too many missing values. To compensate this loss, we define a compensation factor f_{comp} as the sum of the missing items in two underlying time series.
We also adopt the entropy-based evaluation of clustering quality described in Chen et. al's paper [] in finding similar pages. The similarity aware entropy score calculates the entropy of a bunch of time series. Since we've already got the pairwise similarity, we iteratively search for page with the most similar rank series and compute the entropy score until the score exceeds a certain threshold. This is obviously prior to k nearest neighbors because it gets all pages whose rank series are similar enough with that of this page.

D.2 Further Exploring Semantics

Pagelink network

D.3 Implementation

The page view statistics for Wikimedia projects maintains raw page access records for all Wikipedia projects in all languages. We have collected 14-month English page view statistics dataset (from Jun. 1st to Oct. 27th in 2011 and from January to September in 2014) and generated daily top-1000 page views in a mysql data archive. Visualization only shows top-50 page views while page with similar rank series are fetched among the top-1000 pages. We also collected page links dataset via Wikipedia page link APIs and maintained its persistence via Neo4j, a graph database.